VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

ROZPOZNÁVÁNÍ HISTORICKÝCH TEXTŮ POMOCÍ HLUBOKÝCH NEURONOVÝCH SÍTÍ

CONVOLUTIONAL NETWORKS FOR HISTORIC TEXT RECOGNITION

SEMESTRÁLNÍ PROJEKT
TERM PROJECT

AUTOR PRÁCE **Bc. MARTIN KIŠŠ**
AUTHOR

VEDOUCÍ PRÁCE **Ing. MICHAL HRADIŠ, Ph.D.**
SUPERVISOR

BRNO 2018

Abstrakt

Do tohoto odstavce bude zapsán výtah (abstrakt) práce v českém (slovenském) jazyce.

Abstract

Do tohoto odstavce bude zapsán výtah (abstrakt) práce v anglickém jazyce.

Klíčová slova

Sem budou zapsána jednotlivá klíčová slova v českém (slovenském) jazyce, oddělená čárkami.

Keywords

Sem budou zapsána jednotlivá klíčová slova v anglickém jazyce, oddělená čárkami.

Citace

KIŠŠ, Martin. *Rozpoznávání historických textů pomocí hlubokých neuronových sítí*. Brno, 2018. Semestrální projekt. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Michal Hradiš, Ph.D.

Rozpoznávání historických textů pomocí hlubokých neuronových sítí

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením pana X... Další informace mi poskytli... Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Martin Kišš
3. ledna 2018

Poděkování

V této sekci je možno uvést poděkování vedoucímu práce a těm, kteří poskytli odbornou pomoc (externí zadavatel, konzultant, apod.).

Obsah

Kapitola 1

Úvod

Rozpoznávání znaků (anglicky Optical Character Recognition, zkráceně OCR), respektive celých textů, je proces, ve kterém se převádí text z obrazové formy do sekvence znaků, které mohou být dále zpracovány. Vzhledem ke stále větší integraci počítačových systémů do běžného života, může tato úloha nacházet široké uplatnění. Například se může jednat o součást systému řízení autonomního vozidla, které díky rozpoznávání textů je schopné číst dopravní značení. Dále může být součástí aplikací zabývající se překladem z různých jazyků, kdy se překládaný text zadává pomocí kamery (například v chytrém mobilním telefonu). Využití je také možné při vyhledávání obrázků, kdy se díky rozpoznání textů může získat z obrázku více informací, a podobně.

V rámci počítačového vidění se jedná o problém, který je spojen s různými úskalími, která mohou převod z obrazové formy ztížit. Jedná se především o data (např. fotografie), která jsou zašuměná, v různých částech obrazu obsahují různou intenzitu světla, nebo například text může být natočen, zkreslen perspektivou a tak dále.

Cílem této práce je vytvořit nástroj, který bude schopen rozpoznávat text z dokumentů, které vznikly v dřívějších dobách. Jedná se především o text z období středověku, kdy se objevují texty psané ručně a také texty vytvořené pomocí prvních metod knihtisku.

V kapitole 2 budou popsány techniky, které se v současnosti používají k rozpoznávání textu. Dalším obsahem této kapitoly také bude stručný popis vývoje historických písem. Kapitola 3 bude obsahovat zvolenou metodu, která bude v rámci práce implementována a následně vyhodnocena. Další kapitola se bude věnovat popisu datových sad, které jsou pro tuto práci vhodné na otestování. V předposlední kapitole bude popsán současný stav práce a poslední kapitola bude obsahovat závěrečné shrnutí.

Kapitola 2

Rozpoznávání textu

Jak již bylo zmíněno v úvodu, tato kapitola se věnuje přehledu současných řešení rozpoznávání textu. Jedná se výhradně o metody, které jsou založeny na dopředných neuronových sítí, často využívajících také konvoluční vrstvy. Některá takováto řešení budou postupně představena v rámci první části této kapitoly. Ve druhé části pak bude stručně nastíněn vývoj historických písem.

Konvoluční neuronové sítě se začaly hojně využívat ke zpracování obrazových informací po úspěchu z roku 2012, kdy *Krizhevsky a spol.* [2] byli schopni natrénovat hlubokou konvoluční neuronovou síť, jejíž úspěšnost byla lepší, než ostatní dosud běžně používané metody pro klasifikaci obrázků.

2.1 Současné metody založené na neuronových sítích

V současné době se vývoj v oblasti OCR soustředí především na rozpoznávání textu, který se nachází v nějakém prostředí (anglicky se jedná o takzvaný *scene text*). To, že se text nachází v libovolném prostředí, zvyšuje náročnost samotného rozpoznávání, protože výsledný systém by si měl poradit s jakýmkoliv okolím textu a vždy extrahovat informaci pouze o textu samotném, přičemž tento text může nabývat různých podob. Text ve scéně může mít různé styly písma, být zkreslený perspektivou, jiné objekty jej můžou z části překrývat. Dále může mít obrázek na různých místech různé intenzity světla, případně jednotlivé znaky vůbec nemusí být v jedné linii, ale mohou opisovat různé křivky.

Jedním z možných přístupů k rozpoznávání slov ve scéně, je založen na použití histogramu gradientů (anglicky *Histogram of Gradients*, zkráceně *HOG*) [6]. Metoda spočívá ve výpočtu histogramů gradientů ve všech možných výřezech o rozměru 3x3 v původním obrázku. Tímto vznikne mapa histogramů gradientů, která reprezentuje původní vstupní obrázek. Následně je vypočítán průměr všech histogramů, jež se nacházejí ve stejném sloupci. Tímto vznikne vektor gradientů reprezentující jednotlivé sloupce a tento vektor je použit pro rozpoznání daného slova. Pro samotné rozpoznávání slov se zde používá rekurentní neuronová síť (RNN) s použitím obousměrných LSTM (Long Short-Term Memory). Tato síť určuje pro každý vektor pravděpodobnost každého možného znaku.

Pro získání výsledného slova je použita technika CTC (Conectionist Temporal Classification), která vyhodnotí výsledné pravděpodobnosti. Toto vyhodnocení může probíhat ve dvou režimech. První možností, jež byla zároveň v tomto případě použita, je, že CTC v každém momentě, kdy rekurentní neuronová síť dala na výstup pravděpodobnosti znaků, vybere ten s nejvyšší pravděpodobností. Druhou možností je, že se vyhodnotí pravděpodob-

nosti vůči dodanému slovníku. Varianta se slovníkem je nicméně časově náročnější, protože při vyhodnocení se musí vyhodnotit velký počet pravděpodobností. [3]

Rekurentní neuronové sítě jsou obecně považovány za lepší v oblasti vyhodnocování sekvencí, protože se dokážou určitým způsobem naučit různé závislosti v průběhu celé sekvence. Druhou výhodou je, že výstup rekurentní neuronové sítě může být potenciálně nekonečný.

Podobný princip pro rozpoznávání textu, jako výše zmíněný, použili také *Shi a spol.* [3]. V této metodě je však namísto výpočtu histogramu gradientů využito konvolučních vrstev, které získají z obrázku požadované příznaky. Také v tomto případě jsou výsledné pravděpodobnosti zpracovány metodou CTC, avšak zde jsou použity obě možnosti této techniky a jsou navzájem porovnány. Z dosažených výsledků je patrné, že varianta se slovníkem dosahuje vyšších úspěšností.

Jak již bylo zmíněno výše, při rozpoznávání textu ve scéně může být text různě deformován. Jedním z uvedených problémů byl, že jednotlivé znaky textu nemusí spočívat v jedné linii. K řešení tohoto problému využili *Shi a spol.* ve své další práci [4] speciální neuronovou síť zvanou Spatial Transformer Network. Tento typ sítě umožňuje transformovat vstupní obrázek pomocí afinních transformací do požadovaného tvaru. S touto sítí je tedy možné například převést vstupní obrázek, na kteérm se nacházejí písmena tvořící oblouk, na obrázek, kde jsou ta samá písmena relativně v jedné linii. [1]

Spatial Transformer Network (zkráceně *STN*) je modul, který může být použit v rámci jiné neuronové sítě, díky kterému je možné aplikovat na vstup afinní transformace, které se síť, obsažená v tomto modulu, naučí. Celkově se celý modul sestává ze tří částí:

- lokalizační sítě,
- generátoru matice,
- vzorkovače (sampler).

Lokalizační síť má za úkol zpracovat vstupní obrázek a najít takové parametry transformace, které odpovídají požadovanému výstupu. Pro rozpoznávání zakřiveného textu mohou tyto body například reprezentovat ohraničení daného textu [4]. Výstupem této sítě je tady množina hodnot reprezentující transformaci, která bude později aplikována. Generátor matice převezme výstup lokalizační sítě a vytvoří z něj právě výslednou matici transformace. Následně je tato transformace předána sampleru a ten vytvoří na základě této matice a původního vstupního obrázku nový obrázek, který by se měl co nejvíce blížit požadovaných vlastnostem.

Další možností pro rozpoznávání textu je metoda, kterou použili *Springmann a spol.* [5], kteří použili pouze rekurentní neuronovou síť. Vstupem této sítě jsou jednotlivé sloupce vstupního obrázku, ty jsou zpracovány sítí a výstupem jsou jednotlivé znaky. Tato metoda byla použita na řádky historických textů, tedy ne na text ve scéně.

2.2 Historická písma

Písmo, které je používané západní civilizací, tedy latinské písmo (latinka), vzniklo přibližně v 7. století př. n. l odvození z řeckého písma. Postupně vznikalo několik druhů tohoto písma, které se navzájem lišily zkosením, zdobností a podobně. Velká různorodost těchto písem znesnadňovala jejich čtení. Díky tomu byla snaha písmo sjednotit, což vyústilo ve vytvoření několika málo standardních písem.

V rámci druhé poloviny 15. století se středověké latinské písmo rozdělilo do dvou hlavních směrů. Jednalo se o písmo novogotické, které navazovalo na pozdější písma středověká (gotická), a humanistické, které vycházelo z antických a raně středověkých písem. Během novověku spolu obě větve soupeřily, přičemž humanistické písmo nakonec zvítězilo nad novogotickým, které se definitivně přestalo používat v první polovině 20. století.

Kapitola 3

Návrh řešení

Tato kapitola obsahuje popis navrženého řešení pro rozpoznávání historických textů.

Hlavní částí navrženého řešení je rekurentní neuronová síť, která bude schopna zpracovávat řádky textu a produkovat výstup v podobě textové informace. Vstupem do této sítě by měl být výřez z původního obrázku, který obsahuje jeden následující znak celého textu. Součástí této neuronové sítě by také měla být Spatial Transformer Network, kterou lze použít jako jednu z vrstev v rámci komplexnější sítě. Tato vrstva bude mít za úkol přesně lokalizovat v daném výřezu daný znak a ten bude následně pomocí série konvolučních a max-pooling vrstev zpracován tak, aby následné LSTM (případně GRU) vrstvy byly schopné tento znak rozpoznat. Výstupem této sítě bude informace o zpracovaném znaku a také přibližná pozice následujícího znaku. Díky tomuto bude možné zpracovat celý řádek textu najednou.

Popsaná rekurentní neuronová síť bude trénována na datové sadě, která bude vygenerována pomocí implementovaného generátoru historických textů, jenž je popsán v kapitole 5.1. Natrénovaná síť bude následně otestována jednak pomocí vygenerované testovací sady a také pomocí některých reálných datových sad, jež jsou popsány v kapitole 4.

Kapitola 4

Datové sady

Informace ohledně existujících datových sad, jejich stručný popis, atd.

Kapitola 5

Stav řešení

V této kapitole bude popsán stav dosavadní práce. Nejprve zde bude popsán implementovaný generátor historických textů. Motivací k jeho vytvoření bylo, že se nepodařilo najít vhodnou datovou sadu, která by obsahovala naskenované historické text a jejich přepis (tzv. ground truth). Výhodou takto implementovaného generátoru je také množství dat, která mohou být vygenerována, a spolu s tím také možnost přesné lokalizace všech písmen. Ve druhé části této kapitoly bude popsána vygenerovaná datová sada a neuronová síť, která byla na této sadě natrénována. Budou zde také zhodnoceny výsledky, které síť dosáhla.

5.1 Generátor textů

Součástí této práce je generátor umělých textů. Tento generátor vytváří syntetické obrázky takové, aby co nejvíce vypadaly jako originální texty naskenované z historických dokumentů. Ke generování obrázků, které vypadají co nejvěrohodněji, je zapotřebí několika zdrojů. Prvním zdrojem jsou bezpochyby vhodné fonty, které v maximální možné míře odpovídají dobovým písmům. Druhým zdrojem jsou obrázky (textury), které mají vzhled starého papíru a na něž budou výsledné vygenerované texty naneseny.

Aby vygenerované obrázky co nejvíce odpovídaly reálným textům, je na původní vysázený text aplikováno několik takzvaných efektů. Tyto efekty mají za úkol upravovat originální vysázený text tak, aby se co nejvíce přiblížil vzhledu historických textů. Tyto efekty budou popsány v následující části 5.1.1.

K tomu, aby efekty neupravovaly vstupní obrázek pokaždé stejným způsobem, je v generátoru několikrát použito generování náhodných čísel. Aby se daly jednoduše ovládat minimální a maximální hodnoty při tomto generování, je jediným parametrem při spuštění konfigurační soubor, který obsahuje mimo jiné právě tyto minimální a maximální hodnoty pro jednotlivé efekty. Dalšími parametry, které jsou uloženy v rámci konfiguračního souboru, jsou například rozměry výsledného obrázku, adresáře s fonty a podobně. Ukázku celého konfiguračního souboru je možné nalézt v příloze.

K vysázení daného textu do obrázku byla použita knihovna freetype ve verzi pro python [1]. Tato knihovna umožňuje vysázení vlastního textu na poměrně nízké úrovni a je tedy možné během sázení uchovávat užitečné informace, jako například již zmíněné pozice jednotlivých znaků.

[1] https://pypi.python.org/pypi/freetype-py/

5.1.1 Efekty

Jak již bylo zmíněno výše, efekty slouží k úpravě vysázeného textu tak, aby výsledek co nejvíce vypadal, jako reálný historický text. Všechny efekty jsou postupně aplikovány na sebe tak, až vznikne výsledný obrázek.

Jakmile je text vysázen a celý obrázek rozšířen na požadovanou velikost, vzniká kolem původního textu množství volné plochy, která se běžně v dokumentech nevyskytuje. Pro zaplnění tohoto místa je vygenerován další text, který se umístí nad a pod tento text, respektive nalevo a napravo od něj. Tento vytvořený okolní text je náhodně generován z ostatních slov, která se nacházejí ve vstupním textu. Výsledný text, jenž je umístěn nad a pod původní text, je odsazen o hodnotu `LineSpace`, která je definována v konfiguračním souboru. Text napravo a nalevo je obdobně odsazen o hodnotu `WordSpace` a zároveň je vertikálně umístěn tak, aby základní dotažnice obou textů byly navzájem ve stejné výšce.

Dalším efektem v posloupnosti, která je aplikována na obrázek, je efekt, jenž imituje nedokonalosti při historickém tisku. Tento efekt aplikuje na původní obrázek dva typy nedokonalostí. Nejprve je vygenerována matice hodnot v rozmezí 0 až 1, přičemž ke generování se používá knihovna opensimplex [2], která vytváří náhodné souvislé mapy. Tato mapa poté slouží k imitaci nerovnoměrnosti při tisku, kdy se některé části písmen nemusí dobře otisknou na papír a tak je zde text nepatrně bledší. Druhým typem nedokonalosti jsou defekty písmen, kdy se určitá malá část písmene neotiskne vůbec. Toto může být způsobeno například poškozením dané litery, případně nečistotou papíru. Tato nedokonalost je vytvořena tak, že se do matice, která má shodné rozměry jako původní obrázek, náhodně umístí několik náhodných hodnot v rozsahu 0 až 1. Následně se provede Gaussovské rozostření, které dodá těmto bodů plynulejší přechod do zbytku matice a zároveň je nepatrně zvětší. Nakonec jsou hodnoty alfa kanálu původního obrázku vynásobeny hodnotami obou těchto matic.

Následně je s určitou pravděpodobností aplikován efekt, který způsobí, že se ve výsledném obrázku objeví druhý text, který je převrácen přes vertikální osu a zároveň je velice průsvitný. Toto představuje situaci, kdy na se na danou stránku otiskne text z následující strany, případně přes list papíru prosvítá text, který se nachází na druhé straně. Na takto vygenerovaný text se také aplikuje předchozí efekt, tedy efekt, který napodobuje nedokonalosti při tištění. Tento výsledný obrázek je nakonec nanesen na texturu papíru, jež byla pro daný text náhodně zvolena.

Po aplikaci výše zmíněných efektů, je výsledný obrázek nanesen na texturu, na které již může být nanesen prosvítající, případně otisknutý text tak, jako je popsáno v předchozím odstavci. Následně je na tento obrázek aplikován ještě poslední efekt, který způsobí, že některé části písmen vypadají jako rozpité. Tohoto efektu se dosáhne tak, že je vygenerována náhodná mapa, obdobně jako u efektu nedokonalosti tisku, a také se vytvoří nový obrázek, který vznikne aplikací Gaussovského rozostření na původní obrázek. Následně jsou oba obrázky prolnuty s tím, že v daném bodě je výsledná intenzita je rovna součtu hodnot v původním a rozostřeném obrázku, přičemž jedna z hodnot je vynásobena hodnotou ve stejném bodě ve vygenerované mapě a druhá z hodnot je vynásobena hodnotou, která vznikla jako jedna mínus hodnota v mapě.

5.2 Klasifikace umělých textů

Cílem experimentování s konvoluční neuronovou sítí bylo natrénovat ji na klasifikaci slov. Pro síť byla použita architektura, která vychází z klasifikační sítě publikované v [2]. V

[2] https://pypi.python.org/pypi/opensimplex/0.1

tabulce 5.1 jsou popsány jednotlivé vrstvy použité ve výsledné síti, která byla upravena oproti síti z [2] tak, aby rozměry její vrstev lépe odpovídaly velikosti dat.

Typ vrstvy	**Parametry vrstvy**	**Velikost výstupu**
Konvoluční vrstva	16 konvolučních jader	16 x 256 x 128
Max-pooling vrstva	2 x 2 pixely	16 x 128 x 64
Plně propojená vrstva	1024 neuronů	1024
Dropout vrstva	50 %	1024

Tabulka 5.1: Popis vrstev použitých v klasifikační síti.

5.2.1 Datová sada

Z generátoru, který byl popsán v části 5.1, byla vygenerována datová sada, která čítala ..... obrázků. Z tohoto počtu je použito ..... pro trénování sítě, která je popsána výše, a zbylých ..... obrázků je použito na testování. Jako vstupní text bylo zvoleno 300 nejčastějších anglických slov a každé slovo tedy bylo použito ..... krát. Při generování bylo použito ... různých fontů a ... textur papíru.

5.2.2 Výsledky experimentu

Vyhodnocení

Kapitola 6

Závěr

Shrnutí

Literatura

[1] Jaderberg, M.; Simonyan, K.; Zisserman, A.; aj.: *Spatial Transformer Network*. Únor 2016, [Online; navštíveno 19.12.2017].
URL https://arxiv.org/abs/1506.02025

[2] Krizhevsky, A.; Sutskever, I.; Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, editace F. Pereira; C. J. C. Burges; L. Bottou; K. Q. Weinberger, Curran Associates, Inc., 2012, s. 1097–1105.

[3] Shi, B.; Bai, X.; Yao, C.: *An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition*. Prosinec 2016, [Online; navštíveno 19.12.2017].
URL http://ieeexplore.ieee.org/abstract/document/7801919/

[4] Shi, B.; Wang, X.; Lyu, P.; aj.: *Robust Scene Text Recognition with Automatic Rectification*. 2016, [Online; navštíveno 19.12.2017].
URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Shi_Robust_Scene_Text_CVPR_2016_paper.html

[5] Springmann, U.; Lüdeling, A.: *OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus*. Únor 2017, [Online; navštíveno 19.12.2017].
URL https://arxiv.org/abs/1608.02153

[6] Su, B.; Lu, S.: *Accurate recognition of words in scenes without character segmentation using recurrent neural network*. Říjen 2016, [Online; navštíveno 19.12.2017].
URL http://www.sciencedirect.com/science/article/pii/S0031320316303314

Příloha A

Konfigurační soubor

Ukázka konfiguračního souboru pro generátor umělých historických textů:

```
[Common]
Input: input_file.txt
Outputs: Outputs/
Backgrounds: Backgrounds/
Fonts: Fonts/
Annotations: False
Words: True
TrainRatio: 0.8

[OutputSize]
Width: 256
Height: 128
LineHeight: 42

[SurroundingText]
LineSpace: 10
WordSpace: 15

[BackText]
Probability: 0.75
MinAlpha: 0.05
MaxAlpha: 0.15

[PrintingImperfections]
MinCoef: 0.001
MaxCoef: 0.05
MaxColor: 128
SubtractMin: True
MaxBlobs: 50
MinBlobColor: 64
MaxBlobColor: 255

[Blurring]
```

```
MinSigmaX: 0.5
MaxSigmaX: 2.0
MinCoef: 0.01
MaxCoef: 0.1
MaxColor: 255
```